WebTribe: Dynamic Community Analysis from Online Forums

Damien Leprovost[1], Lylia Abrouk[1], and David Gross-Amblard[2]

[1] Le2i CNRS Lab, University of Bourgogne, Dijon, France
firstname.lastname@u-bourgogne.fr
[2] IRISA, University of Rennes I, France

Abstract. In this demonstration we present WebTribe, a tool for community discovery based on the analysis of large discussion forums or e-mail repositories[3]. In this tool, communications are tracked in real time, analyzed according to a reference ontology, and a summary of users' activity is built in an incremental way. The demonstration will illustrate how communities are identified and updated depending on the semantics and structure of communications between users.

1 Introduction

Discussion forums constitute a well-known advertising tool for companies, as they attract existing and potential customers on the company's website, give product insights, and show the company openness and activity. In this context, the *community manager* is an emerging role in such companies. Typically, the community manager, aside the traditional task of moderating forums and managing topics, has to monitor the forum activity, report on existing sub-communities, identify expert users and opinion leaders for specific targeting (advertising, special offers, ...). But due to the exploding rate of forum contributions, monitoring tools are needed to assist the manager.

In this demonstration we will present WebTribe, a system that allows community managers to perform these tasks on various kind of forums or public e-mails archives in a scalable and incremental way. Our model encompasses every type of user communications (forums, tweets, emails, ...), as soon as a specific wrapper is provided (we give such a wrapper for a specific healthcare company forum). Several analysis axes can be considered in forum analysis: users connections and posting rates, citations (replies) between users, and post content. Existing methods usually rely for the latter on term frequencies, a method that allows to give a rough overview of the forum activity. In WebTribe, we enable the community manager to be active, by giving a controlled term vocabulary in the form of a target ontology. It also allows reasoning within the ontology: a user posting terms (concepts) such as *ventricle*, *aorta* or *vena cava* will be identified as a *heart* expert, while this term never appears explicitly in the user's posts.

[3] An earlier version of this demo has already been presented at the French conference BDA 2011, which has informal proceedings and does not retain any copyright.

Concept analysis allows a real-time interpretation of the evolution of communities. This demonstration proposal is organized as follows. After briefly presenting the related work, we present our model (Section 2) and detail the architecture of the WebTribe system (Section 3) along with the scenario of our demonstration.

Related Work The importance of comment activity on blogs was the subject of several studies [5]. Previous works have focused on highlighting the structure of discussions within new articles, in order to determine popular topics, conflicts of opinion [7, 3, 1], or relational implications between users [2, 6]. In these works, ontologies are not used to structure the vocabulary or refine the analysis. Dedicated ontologies like SIOC[4] exist for structuring forums. Our approach is complementary, as it allows a community manager to analyze external forums with a specific ontology (say a brand product), different from the forum's SIOC. Moreover, there are still numerous forums without such SIOC structurations. The model underlying this demo was detailed in our previous work [4].

2 Model and Architecture

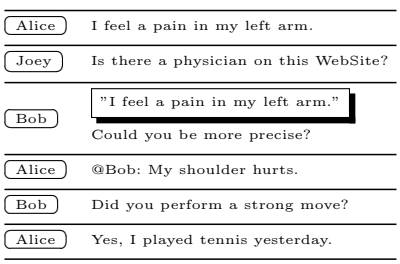

Fig. 1. Posts in a forum

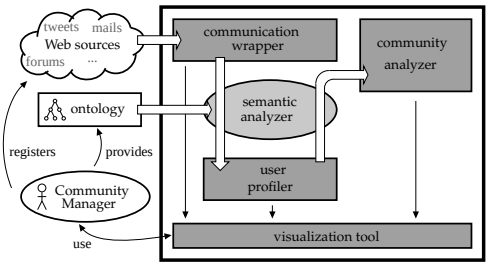

Fig. 2. WebTribe architecture

Figure 1 shows an example of communications in a healthcare forum, where Alice, Joey and Bob discuss. In order to finely define the axis of the forum semantic analysis, we rely on a domain or generic ontology, as uses in Figure 3, which describes concepts with their subconcept relation (we restrict our attention to structural relations like is-a, part-of, sort-of, etc.). Choosing a target ontology enables a flexible forum exploration: a generic ontology like WordNet for a forum overview, a specific, e.g. brand product ontology for a specific tracking of topics.

Given a post, we identify its author (according to the forum API or syntactic rules of the Web page). The set of concepts occurring in the post is computed by stemming the post and removing stop-words, and by comparison with the (stemmed terms of the) ontology (for example, terms `arm` and `shoulder` in Figure 1 are identified as relevant concepts). During posts analysis, the running *user profile* for each user is computed, as the sum of concepts occurrences in the ontology. Figure 3 shows the profile of a user who used once the concept

[4] `http://www.sioc-project.org/`, 2012.

`shoulder`, 16 times `humerus` and 8 times `biceps`. Interpreting such profiles can be tedious, due to the huge potential number of concepts. In order to overcome this difficulty, we build a *user abstract* by saturating the following rules:

Relevance If a concept occurrence, relatively to the other concepts, is smaller than the relevance threshold $\delta_{relevance}$, the concept is discarded in the abstract. This limits the impact of terms used occasionally, and favor long-term interests.

Coverage If almost all subconcepts of a concept c are covered (non-zero occurrence), the concept c itself receives the average occurrence. The fraction of covered concepts required is controlled by the $\delta_{coverage}$ threshold. This models the fact that a user, talking significantly about `biceps`, `humerus` and `triceps`, should indeed be considered as talking about `arm`, with the corresponding strength.

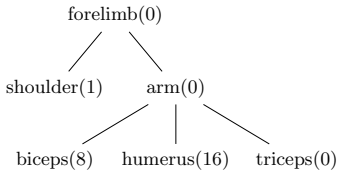

Fig. 3. User profile

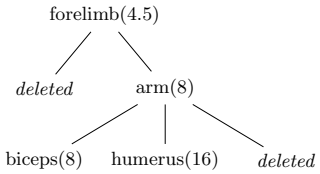

Fig. 4. User abstract

Observe that we do not rely on a tf-idf computation for concepts detection because we want to perform generalization. Choosing the value of these thresholds depends on the ontology and the forum pace, and is managed for now by a manual tuning. As an example recall the profile of Figure 3. For $\delta_{coverage} = 0.66$ and $\delta_{relevance} = 1/24$, the resulting *abstract* appears in Figure 4.

Finally, the *forum abstract* is computed in a similar way: we sum the abstract of all users and only apply the relevance threshold. Communities are then identified by the top k concepts with the largest occurrence (each community is identified by a unique concept). Users may belong to several communities, proportionally to their concepts occurrence in their abstract. For example, if *arm* and *shoulder* turn out to be the two communities of the system (the top 2 concepts), the user of Figure 4 belongs to the first community with score 8, and does not belong to the second.

We enrich the previous analysis by taking into account the context of communication. there are several technical or textual conventions for answering a given post. For emails or tweets the user who is answered to is explicitly given. For purely web systems, classical patterns are to start the answer to user u with "@u", or to cite the answered message. In Figure 1, the first post of Bob explicitely cites Alice's post, hence the `arm` concept is propagated in Bob's post. The second post of Bob is an implicit answer to the previous post: we then propagate the previous `shoulder` concept into Bob's post. Figure 2 presents the general WebTribe architecture.

3 Demo Scenario

Our 10mn demo considers a community manager taking over the health section of the USA Today forum[5]. We will illustrate the following functionalities, available as a video at `http://www.damien-leprovost.fr/webtribe`:

1. **Source registration, Ontology selection** The manager selects the forum URL and a target ontology (as an OWL file or a subtree of WordNet, given a root concept).
2. **Visualization** During the entire demonstration, the whole activity can be monitored. For example, as the target forum is analyzed on the fly, a specific window allows seeing the forum with ontology concepts highlighted. Given a community, both its main topic and users can be displayed. For a user, her/his main manipulated topics and possible related communities are listed.
3. **Forum health status** A global indicator of the community is given, that measures its global health: number of users, covered topics, activity rate, ...
4. **Alert system** A simple alert language allows to monitor the activity at the post / user / community level, to warn the community manager of any interesting event (for example, the first use of the name of a disease).
5. **Multiple forum analysis** Finally, the system can also perform the analysis of several sources, with the same workflow. It allows comparisons, in order to detect similar communities, potential new users, and login equivalences.

References

1. Amer-Yahia, S., Lakshmanan, L., Yu, C.: Socialscope: Enabling information discovery on social content sites. In: Conference on Innovative Data Systems Research (CIDR) (Sep 2009), `http://arxiv.org/abs/0909.2058`
2. De Choudhury, M., Mason, W.A., Hofman, J.M., Watts, D.J.: Inferring relevant social networks from interpersonal communication. In: International conference on World wide web (WWW). pp. 301–310. ACM, New York, NY, USA (2010)
3. Gloor, P.A., Zhao, Y.: Analyzing actors and their discussion topics by semantic social network analysis. In: Conference on Information Visualization. pp. 130–135 (2006)
4. Leprovost, D., Abrouk, L., Gross-Amblard, D.: Discovering implicit communities in web forums through ontologies. Web Intelligence and Agent Systems: An International Journal 10, 93–103 (2011)
5. Menchen-Trevino, E.: Blogger motivations: Power, pull, and positive feedback. Internet Research 6.0 (2005)
6. Mitrović, M., Paltoglou, G., Tadić, B.: Quantitative analysis of bloggers' collective behavior powered by emotions. Journal of Statistical Mechanics: Theory and Experiment 2011(02), P02005 (2011)
7. Schuth, A., Marx, M., de Rijke, M.: Extracting the discussion structure in comments on news-articles. In: ACM international workshop on Web information and data management (WIDM). pp. 97–104. ACM, New York, NY, USA (2007)

[5] `http://yourlife.usatoday.com/health/`, 2011.